Supporting Text

All supplementary files are available at http://web.mit.edu/leonid/modules. All clusters with functional annotation can be found at http://web.mit.edu/leonid/modules/allclusters.html.

Cluster Search and Processing Methods

The search for clusters in protein-protein interaction network is performed in three steps: (*i*) the graph is processed to minimize the number of proteins and interactions that would give statistically insignificant clusters; (*ii*) three methods are used to identify clusters on this graph; and (*iii*) the obtained clusters are processed to remove redundant clusters and merge overlapping ones.

Graph Processing

The power-law degree distribution of the protein-interaction network means that there are several proteins with a large (100–200) number of interactions. These proteins are likely to lead to formation of star-like clusters with low statistical significance. All proteins that have >30 interactions are removed from the graph to avoid finding such spurious clusters.

Cluster Search

Enumeration of Complete Subgraphs. The search for complete subgraphs (cliques) starts from the smallest statistically significant size, which is 4 in this graph, and goes up in size one by one until all cliques are found.

To find cliques with size 4, all pairs of edges (6,500 × 6,500) are picked successively. For every pair *A–B* and *C–D* we check whether there are edges between *A* and *C*, *A* and *D*, *B* and *C*, and *B* and *D*. If all of these edges are present, *ABCD* is a clique.

For every found clique *ABCD* we pick all known proteins successively. For every picked protein *E*, if all of the interactions *E–A*, *E–B*, *E–C*, and *E-D* are known, then *ABCDE* is a clique with size 5. This search is continued for cliques of size 6, 7, and larger until the biggest clique is found. For our protein-interaction network the largest such clique has size 14.

These results, however, include many redundant cliques, because, for example, the clique with size 14 contains 14 cliques with size 13. To find all nonredundant subgraphs, we mark all proteins comprising the clique of size 14, and out of all subgraphs of size 13 we pick those that have at least one protein other than marked. After all redundant thirteens are removed, we proceed to remove redundant twelves. This process is continued for cliques with size 11, 10, and smaller until no redundant cliques are left. In total, 41 nonredundant cliques with sizes 4–14 are found.

Superparamagnetic Clustering (SPC). SPC uses physical properties of an inhomogeneous ferromagnetic model to find tightly connected clusters on a large graph (1,2). In this approach, every node on the graph is assigned a Potts spin variable $S_i = 1, 2, \ldots, q$ (3). The value of this spin variable performs thermal fluctuation, which is determined by the temperature T and the spin values on the neighboring nodes. Energetically, two nodes connected by an edge are favored to have the same spin value, so the spin at each node tends to align itself with the majority of its

neighbors. Therefore, when such a Potts spin system reaches equilibrium for a given temperature T, high correlation between fluctuating S_i and S_j at nodes i and j would indicate that nodes i and j belong to the same cluster.

The protein-interaction network is represented by a graph where every pair of interacting proteins is an edge of length 1. The simulations are run for temperatures ranging from 0 to 1 in units of the coupling strength. The network splits to monomers at temperatures between 0.7 and 0.8, whereas larger clusters exist for temperatures between 0.1 and 0.7.

Clusters are recorded at all values of temperature. The overlapping clusters are then merged and redundant ones are removed by the procedure described below.

Monte Carlo (MC) Simulation. This method is used to find a tight subgraph of a predetermined number of nodes M. At time $t = 0$ a random set of M nodes is selected. For each pair of nodes i,j from this set, the shortest path L_{ij} between i and j on the graph is calculated. Denote the sum of all shortest paths L_{ij} for this set as L_0. At every time step one of M nodes is picked at random, and one node is picked at random out of all its neighbors. The new sum of all shortest paths, L_1, is calculated if the original node were to be replaced by this neighbor. If $L_1 < L_0$, the replacement takes place with probability 1, if $L_1 > L_0$, the replacement takes place with probability $\exp(-(L_1 - L_0)/T)$, where T is the effective temperature. Every tenth time step an attempt is made to replace one of the nodes from the current set with a node that has no edges to the current set to avoid getting caught in an isolated disconnected subgraph. This process is repeated until the original set converges to a complete subgraph, or for a predetermined number of steps, after which the tightest subgraph (the subgraph corresponding to the smallest L_0) is recorded.

For every cluster size there is an optimal temperature that gives the fastest convergence to the tightest subgraph. Fig. 6 shows the dependence of time to find a clique with size 7 in MC steps per site as a function of temperature T. The required time increases sharply as the temperature goes to zero, but has a relatively wide plateau in the region $3 < T < 7$. Our simulations at other sizes suggest that the choice of temperature at $T \approx M$ would be safe for any cluster size M within the range of sizes we used.

As a modification of the MC method, we run simulation starting with a *connected* set of nodes, meaning every node is a neighbor of at least one of the other nodes. At every step when a node is picked, an attempt is made to replace it with *a neighbor of any of the remaining nodes* rather than its neighbor or an arbitrary node on the graph. The replacement is made by the same rules as in the previous paragraph.

The recorded clusters are merged and redundant clusters are removed for both MC methods.

Merging Overlapping Clusters. A simple statistical test shows that nodes which have only one link to a cluster are statistically insignificant. Therefore, before any merging of the overlapping clusters is done, all clusters are "cleaned" of such statistically insignificant members. Merging the overlapping clusters then proceeds as follows. For every cluster A_i we find all clusters A_k that overlap with this cluster by at least one protein. For every such found cluster we calculate the Q value of a possible merged cluster $A_i \cup A_k$. We then record the cluster $A_{best}(i)$, which gives the highest Q value if merged with A_i. After the best match is found for every cluster, every cluster

A_i is replaced by a merged cluster $A_i \cup A_{best}(i)$, unless $Q(A_i \cup A_{best}(i))$ is below a certain threshold value Q_c. This process continues until there are no more overlapping clusters or until merging any of the remaining clusters will make a cluster with Q value lower than Q_c. The Q_c we used was 0.33 for SPC and 0.2 for MC.

Comparison of SPC and MC Results

Clusters obtained by SPC and MC simulation show a reasonably good overlap, but at the same time these two methods complement each other very nicely. Fig. 7 shows clusters found with SPC (red) and MC (blue). Almost one-third of all clusters were found by both methods.

SPC method strongest side is high-Q value clusters with relatively few links with the outside world. The best example is TRAPP complex, a fully connected clique of size 10 with just 7 links with outside proteins. This cluster was perfectly detected by the SPC, whereas MC simulation was able to find smaller pieces of this cluster separately rather than the whole cluster. By contrast, MC simulation is better suited for finding very "outgoing" cliques. The Lsm complex, a clique of size 11, includes three proteins with more interactions outside the complex than inside. This complex was easily found by MC, but was not detected as a stand-alone cluster by SPC.

Rewiring Procedures for Tests of Statistical Significance

The rewiring is performed as follows (4). Two links connecting four different vertices (*A*-*B*, *C*-*D*) are picked at random and rewired such that the new links are *A*-*C* and *B*-*D*, provided that neither of these two new links existed before, or *A*-*D* and *B*-*C* with the same condition. This procedure continues until the fraction of links rewired at least once reaches the desired value. This method guarantees the preservation of the distribution of connectivities.

Rewiring the graph can also serve to investigate the robustness of our methods against false-positive protein-protein interactions and experimental errors. We ran the SPC on a graph altered in the following three ways. First, a certain fraction of links was rewired by using the above procedure. Second, a certain fraction of links was randomly removed from the graph altogether. And third, a certain fraction of links was added to the graph. For every fraction of links altered, we repeated SPC cluster search 10 times. For every cluster we found with the SPC on the original graph, we found the fraction of its proteins recovered on every rewired graph. We then plot the probability to recover 50% and 75% of any original cluster as a function of the fraction of altered links (Fig. 9).

The effect or edges removal or addition on the cluster recovery probability is relatively small. This demonstrates that SPC is robust with respect to a potentially large number of false-positive protein-protein interactions. This also demonstrates the ability of SPC to recover the original clusters when more links are added to the graph as a result of future experiments.

Statistical Significance. If clusters are obtained by some optimization techniques (e.g., MC optimization) that seeks to maximize *m*, given the size of the cluster *n*, then the probability to observe a cluster with no less than *m* connections approximately follows the Fisher-Tippett extreme value distribution (EVD)

$$P_{evd}(m) = \exp(\exp(-\alpha(m - u))), \quad \textbf{[1]}$$

where α and u are parameters of the distribution.

We used this property to estimate statistical significance of clusters. First, we generate 1,000 randomly rewired networks and run MC search on each of them. This way we obtained clusters with m obeying EVD and derived parameters $\alpha(n)$ and $u(n)$ as a function of cluster size n. Second, we analyzed clusters discovered in the real network and computed P_{evd} for each of them by using Eq. **1**. Clusters with $P_{evd} < P_{cutoff} = 10^{-4}$ were said to be statistically significant. Fig. 1*B* presents the distribution of Q and their EVD approximations obtained by using randomly rewired networks together with the clusters discovered in the real network.

We noticed that $\alpha(n)$ and $u(n)$ scale linearly with n

$$\alpha(n) = (a_1 n + a_2)^{-1}; \; u(n) = u_1 n + u_2, \quad \textbf{[2]}$$

allowing computation of P_{evd} for a cluster of any size n. The exact values of parameters we used were $u_1 = 1.19478$, $u_2 = -1.662594$; $a_1 = 0.104079$, $a_2 = -0.135477$. One can also establish P_{cutoff} and then invert these expressions to obtain $Q(n)_{cutoff}$, such that any cluster of n nodes

$$Q(n)_{cutoff} = \frac{m_{cutoff}}{n(n-1)/2} = 2\frac{Ra_1 + u_1}{n-1} + 2\frac{Ra_2 + u_2}{n(n-1)}$$

with $Q > Q(n)_{cutoff}$ is considered to be statistically significant. Due to linear scaling of $\alpha(n)$ and $u(n)$ we get

where $R = -\log(-\log(1 - P_{cutoff})) \approx -\log P_{cutoff}$. Fig. 8 presents $Q(n)_{cutoff}$ for three different values of P_{cutoff}. Note that these values of $Q(n)_{cutoff}$ are not transferable to other graphs because they have been obtained by rewiring this particular network. The methodology, in contrast, can be applied to any graph and can provide estimates of $Q(n)_{cutoff}$.

Comparison with Experiment and Predictions

Comparison with experiment has two goals: (*i*) to verify that our methods correctly identify known experimental complexes (5–8) by using the graph of pairwise interactions, and (*ii*) to identify previously uncharacterized complexes and complex members that were not found in previous experiments.

For the first goal one should notice that complexes and pairwise interactions are being searched for by independent experimental methods, therefore, for many proteins in the experimental complexes there are no known pairwise interactions. Because our prediction methods are based on using the pairwise interactions, one should only pick those complexes for comparison that have similar statistical significance to ours. Using the cutoff $Q = 0.2$ and $P_{evd} < 0.001$, one finds 29 experimental complexes.

For the second goal we will compare our clusters with *all* known experimental complexes. We will identify (*i*) previously uncharacterized complexes, (*ii*) previously uncharacterized membership of proteins with known function, and (*iii*) possible function for

proteins whose complex membership suggests functional correlation.

To choose the best overlap B for a given complex A, consider the probability P that in a graph of N vertices the complex A of size M_A overlaps with a complex B of size M_B by O vertices. If T is the total number of ways to select a set of M_B vertices out of N available on the graph and S is the number of those sets, where O vertices belong to the experimental cluster, then $P = S/T$, where

$$T = \binom{N}{M_B}$$

and S equals the number of ways to select $M_B - O$ vertices out of $N - M_A$ that do not belong to the experimental complex times the number of ways to select O vertices out of M_A that do belong to it. Putting everything together,

$$P = \frac{\binom{M_A}{O}\binom{N-M_A}{M_B-O}}{\binom{N}{M_B}}.$$

The cluster B with the smallest P is the best match for the given complex A.

It is easy to estimate the expectation value that a match as good as a given one occurred accidentally when comparison is run over a large set of clusters by multiplying the found probability P by the size of this set. We consider the match significant if this expectation value is less than 0.01.

We find that *all* experimental complexes satisfying the criteria $Q > 0.2$ and $P_{\text{evd}} > 0.001$ have been identified by at least one of our numerical methods. We report the best match for every experimental complex at http://web.mit.edu/leonid/modules/experimental_identified.html.
The clusters that do not overlap with any known experimental complexes are suggested to be novel complexes or pathways. Such clusters found by our numerical methods are summarized at http://web.mit.edu/leonid/modules/predictions_merged_fa.html. A brief description follows.
–YIP complex (CG0005): One protein has been characterized as YIP1 golgi membrane protein. Others have no clear annotation while sharing some homology with membrane proteins.
–Peroxisomal transport complex (CG0006).
–SNO complex (CG0009): Two proteins (SNZ1 and SNZ2) are members of stationary phase protein family and belong to functional category “Cell rescue, defense, and virulence–Stress response.” Four proteins (SNZ3, SNO1, SNO2, and SNO3) have strong similarity to proteins from this functional category. Unclassified ORF YMR322C is therefore also a possible member of this category.
–Set of four ORFs with no known proteins or function (SP0045). A noteworthy feature is that these are two pairs of adjacent ORFs.
–NN0514: GRX5 is a member of “Cell rescue, defense, and virulence–Stress response” functional category, whereas the other four ORFs have similarity to probable transcription factors.
–NN0519: Four proteins (TRM1, MEP1, EPT1, and GTT1) are involved in cellular transport, YEL017W not classified and possibly belongs to this functional category too.

–NN0631: Five members are clear transcription–mRNA processing category, whereas SMY2 and a nonclassified ORF YPL105C are members of the "Cellular transport" functional category.
–NN0637 autophagy complex. Functional category: "Control of cellular organization."
If a protein that was not found to belong to an experimental complex has a significant number of known pairwise interactions with this complex members, then this protein could be a member of this complex not identified previously in the experiment. We identify several such cases.
CG0003: MI0065 should probably include STE50 protein.
MI0133: possibly includes TIP20 protein (from CG0012) and SFT1 (from CG0016 or SP309).
CG0026: MI0354 possibly includes RAD52 protein.
CG0034, SP0127: BI0359 and CZ0004 possibly include Ser/Thr protein kinase GIN4.
SP0001, NN0522: histone complex MI0048 possibly includes transcription elongation protein SPT6.
MC3936: MI0436 possibly includes TAF145, ADR1 and SPT15.
NN0510: MI0195 possibly has a subcomplex.
NN0516: MI0102 possibly has a subcomplex.
NN0852: MI0096 possibly contains NIP29 and CKS1.
NN0857: MI0404 possibly contains YPL105C.
NN0623, NN1072, NN1284, NN1310, NN1460, NN1467, and NN1625 are new "modules," but each has several proteins that overlap with experimental complexes. NN0623 includes two 40S small ribosomal subunits, RPS28A and RPS28B that are possible members of the Lsm complex.
Finally, we identify four unclassified ORFs that have significant numbers of interactions within our numerically found complexes. This suggests that these proteins have similar function to that of the complex they belong to:
CG0009: YMR322C is possibly a member of "Cell rescue, defense, and virulence–Stress response."
NN0519: YEL117W is possibly involved in cellular transport.
CG0013, SP0424: DAD1 possibly a spindle pole protein.
CG0020: YNR053C is possibly involved in pre-mRNA splicing.

1. Blatt, M., Wiseman, S. & Domany, E. (1996) *Phys. Rev. Lett.* **76,** 3251.
2. Blatt, M., Wiseman, S. & Domany, E. (1997) *Neural Comput.* **9,** 1805.
3. Jain, A. K. & Dubes, R. C. (1998) *Algorithms for Clustering Data* (Prentice–Hall, Englewood Hills, NJ).
4. Maslov, S. & Sneppen, K. (2002) *Science* **296**, 910–913.
5. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002) *Nucleic Acids Res.* **30,** 31–34.
6. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001) *Nucleic Acids Res.* **29,** 242–245.
7. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al.* (2002) *Nature* **415,** 180–183
8. Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002) *Nature* **415,** 141–147.